

## **Appendix A**

### **Statistical Methods Used to Compute Upper Confidence Limits for the Mean Inventory of Contaminants Based on CWID**



## Appendix A

### Statistical Methods Used to Compute Upper Confidence Limits for the Mean Inventory of Contaminants Based on CWID

This analysis computed the summary statistics and upper confidence limits (UCLs) for the concentrations of contaminants at specific release sites.

#### Data issues:

The units of measure used in the CWID database (DOE-ID 2000) were picocuries per gram (pCi/g) for radiological contaminants, and a variety of units for nonradiological contaminants. In particular, the units for nonradiological measurements were either mg/kg, ug/g, ug/kg, ug, su or “units”. The first two of these, mg/kg and ug/g, are basically the same thing, so ug/g was simply renamed to mg/kg. The concentrations in ug/kg were then converted to mg/kg. This amounted to simply multiplying each such concentration by  $1e-03$ .

The contaminants with units given in ug were ytterbium, terbium, and dysprosium, and all were at release site WRRTF-01. Because the soil sampled was not indicated for these contaminants, it was not possible to determine concentrations and these measurements were excluded from the analysis. The measurements in either su or “units” were measurements of pH, and these were also excluded from the analysis.

Methods for handling nondetects (i.e., less than detectable measurements) are discussed by Hertzler, Atwood and Harris (1989). If we denote by DL the detection limit, then it is suggested that nondetects could be replaced in the formulas for the sample mean and sample variance by either 0,  $DL/2$ , or DL. It is noted that if 0 is used, the resulting estimate of the mean is biased low, but the estimate of the standard deviation is biased high. On the other hand, if nondetects are replaced by DL, the estimate of the mean is biased high, but the estimate of the standard deviation is biased low. In order to obtain a UCL which tends to be conservative, we use the estimate of the mean which replaces nondetects with DL and the estimate of the standard deviation where nondetects are replaced with 0. If the concentrations are all detectable, then the estimates are simply the usual sample mean and sample standard deviation.

In some cases where the concentration was less-than-detectable, instead of recording a detection limit in the CWID database, a zero was recorded. When this is the case for all measurements, it is not possible to compute a UCL because no estimate of the variance can be computed. Also, there are cases where only one concentration was measured, and, when this is the case, no estimate of the variance can be computed, so again no UCL can be computed. There are also several cases with radiological measurements in which the data recorded in the CWID database are negative, even though, in reality, concentrations must be non-negative. This often occurs when the concentrations are near or below detection limits. It is not unusual in this instance for the average of background measurements to be greater than some (or all) of the concentration measurements. The background adjustment subtracts the average background from each measurement, and, when the average background is greater than the measurement, the result is a negative value. When computing a UCL, it makes sense to replace a negative estimate of the mean with zero, but the negative measurements still contain useful information about the amount of variability and can be used to estimate the standard deviation and compute a UCL.

Various approaches for computing one-sided UCLs for the mean concentration of a contaminant are discussed by Singh, Singh and Engelhardt (1997). Specifically, if a normal distribution provides an adequate fit to the data, then the standard approach in constructing a UCL based on the Student's t is recommended. Otherwise, a nonparametric UCL based on the Chebychev bound is recommended.

The Student's t approach assumes that the data are independent measurements, each distributed according to a common normal distribution. The assumption of normality can be checked by running a goodness-of-fit test for normality on the data.

The Chebychev approach makes no special parametric assumption such as normality. Instead, it assumes only that the data are independent measurements from a common distribution with an unspecified form. The resulting UCLs tend to be conservative in the sense that they tend to include the true mean with higher than the nominal confidence level.

### Formulas:

Let  $x_1, x_2, \dots, x_n$  be a set of  $n$  measurements resulting from a random sample from a population with mean  $\mu$  and standard deviation  $\sigma$ . Denote the sample mean and sample standard deviation, respectively by

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

and

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

If the population is normally distributed, then a  $100 \times (1 - \alpha)\%$  UCL is of the form

$$\bar{x} + t_{1-\alpha, n-1} \frac{s}{\sqrt{n}} \quad (1)$$

where  $t_{1-\alpha, n-1}$  is the  $100 \times (1 - \alpha)\%$ th percentile of the Student's t distribution with  $n - 1$  degrees of freedom.

For example, if  $\alpha = 0.05$ , then (1) is a 95% upper confidence limit. A common interpretation is that on the average 95% of such limits will lie above the true population mean  $\therefore$ .

Even if the population is not normal, the UCL of equation (1) is often justified by invoking the Central Limit Theorem (CLT) if the number of measurements,  $n$ , is sufficiently large. An often cited guideline for use of the CLT is when  $n > 30$ , in which case the t-value is replaced by the corresponding asymptotic normal value. For example, the standard normal 95<sup>th</sup> percentile is  $z_{0.95} = 1.645$ . Strictly speaking, the CLT requires the use of the population standard deviation,  $\sigma$ , but because this is usually unknown it is common practice to use in its place the sample standard deviation,  $s$ .

The Chebychev inequality has the form

$$P[-k\sigma \leq X - \mu \leq k\sigma] \geq 1 - 1/k^2$$

where  $X$  is a random variable with mean  $\mu$  and standard deviation  $\sigma$  (see, e.g., Bain and Engelhardt [1992, page 76]). If we apply the Chebychev inequality to the arithmetic average  $\bar{X}$  of  $n$  independent random variables, each distributed the same as  $X$ , then we have

$$P\left[-\frac{k\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq \frac{k\sigma}{\sqrt{n}}\right] = 1 - 1/k^2.$$

By equating the right side of the equation to 0.95 and solving for  $k = 4.47$  we have that we are at least 95% confident that

$$-\frac{4.470}{\sqrt{n}} \leq \bar{x} - \mu \leq \frac{4.470}{\sqrt{n}}.$$

If we disregard the right side of this inequality and with a little simple algebra, we are more than 95% confident that

$$\mu \geq \bar{x} - \frac{4.470}{\sqrt{n}} \quad (2)$$

so that the right side of (2) is a conservative 95% UCL for  $\mu$ . Strictly speaking, to apply this limit requires knowing the population standard deviation, but, as suggested by Singh, Singh and Engelhardt (1997), we will use the sample standard deviation as an estimate.

There are two issues to be resolved regarding the use of equations (1) and (2). In particular, it is necessary to perform a goodness-of-fit test for normality, and also, when some of the measurements are below detection limits, it is necessary to replace the estimates with approximations or conservative upper bounds.

Concerning the goodness-of-fit testing, there exist many such tests which are designed to test whether the normal model provides a reasonable fit. The one we are using here is based on the skewness statistic and is designed to be sensitive to deviations from symmetry, a well-known property of the normal distribution. The details of this test are discussed by D'Agostino and Stephens (1986, page 377).

The test is based on a standardized version of the third sample moment

$$m_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n}$$

Specifically, the statistic has the form

$$\sqrt{b_1} = m_3 / m_2^{3/2} \quad (3)$$

where

$$m_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

The null distribution for this statistic is discussed in D'Agostino and Stephens (1986, page 281), and a transformation is suggested which (3) is approximately standard normal when  $n \geq 8$ . This was used to test for normality.

It is possible to obtain a conservative UCL in the case where all measurements are nondetects, but a different rationale is required for the estimate of standard deviation. Suppose, for example, all measurements attempted are below the detection limit. Then, the sample standard deviation cannot be computed exactly because no exact measurements are available. However, it is known that all measurements are between 0 and the detection limit. If the detection limit is the same for all measurements, we denote it by DL, otherwise denote the maximum detection limit by DL. We consider the case where the number of measurement is even. The argument would be similar for an odd number. An upper bound for the sample variance  $s^2$  would be obtained by applying the formula for  $s^2$  with half of the  $x_i$  were set equal to 0 and the other half set equal to DL. For data of this sort, the mean formula would yield  $DL/2$ , and the variance formula would yield

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{(n/2)(0 - DL/2)^2 + (n/2)(DL - DL/2)^2}{n-1} = \frac{n}{n-1} (DL/2)^2$$

and, consequently, we have an upper bound for the sample standard deviation,

$$s \leq \sqrt{\frac{n}{n-1}} \frac{DL}{2} \quad (4)$$

If the number of measurements is odd, it can be shown that instead of (4) an upper bound for the sample standard deviation is given by the slightly different form

$$s \leq \sqrt{\frac{n+1}{n}} \frac{DL}{2} \quad (5)$$

The arithmetic average of the detection limits provides an upper bound for the sample mean when all measurements are nondetects. Thus, in the case of all nondetects, a conservative UCL can be obtained based on these upper bounds.

## REFERENCES

- Bain, Lee J. and Max Engelhardt, 1992, *Introduction to Probability and Mathematical Statistics*, Boston: Duxbury Press, p. 76.
- D'Agostino, R. B., and M. A. Stephens, 1986, *Goodness-of-Fit Techniques*, New York: Marcel Dekker, pp. 281,377.
- DOE-ID, 2000, *CERCLA Waste Inventory Database Report for the Operable Unit 3-13 Waste Disposal Complex*, DOE/ID-10803, Rev. 0, December 2000.
- Hertzler, C. L., C. L. Atwood, and G. A. Harris, 1989, *Current Methods of Handling Less-Than-Detectable Measurements and Detection Limits in Statistical Analysis of Environmental Data*, EGG-SARE-8609, EG&G Idaho Inc., Idaho National Engineering Laboratory, Idaho Falls, Idaho.
- Singh, Ashok K., Anita Singh, and Max Engelhardt, 1997, *The Lognormal Distribution in Environmental Applications*, EPA/600/R-97/006, U.S. Environmental Protection Agency.

**Appendix B**  
**Organic Contaminant Design Inventory**







[illegible]

Chemical	Concentration	Concentration	Concentration
Chlorine	Concentration estimated from the mean weighted average concentrations.	Concentration identified as CHMID or referenced report.	Concentration not expected to be present at site
Chlorine	Concentration identified as CHMID or referenced report.	Concentration not expected to be present at site	Concentration estimated from the mean weighted average concentrations.

Organic compound not typically analyzed and not included on the CLP Organic Compound List. The compound may have been identified on the Appendix list, priority published list, or as a TCE.

1) Printed container and/or container label concentration at 10% based on a Regulatory Analysis and Reassessment of U.S. Environmental Protection Agency Listed Hazardous Waste Numbers for Applicability to the INTEC Liquid Waste System (NCE/CE/96-01213, February 1999).

2) Data based on Table 4-17 of the HHS/EA (OCE/EA-03557, November 1997).

3) Data based on Preliminary Scoping Track 2 Summary Report for Operable Unit 1-03 (EIS-94-10554, April 1993).

4) CPD-6 was combined with CPD-7 from the CMO.

5) CPD-3A was combined with CPD-3 from the CMO.

6) CPD-01A5 was combined with CPD-01A5 from the CMO.

7) CPD-01A5 was combined with CPD-01A5 from the CMO.



